W1409

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-076710
(43)Date of publication of application : 14.03.2003

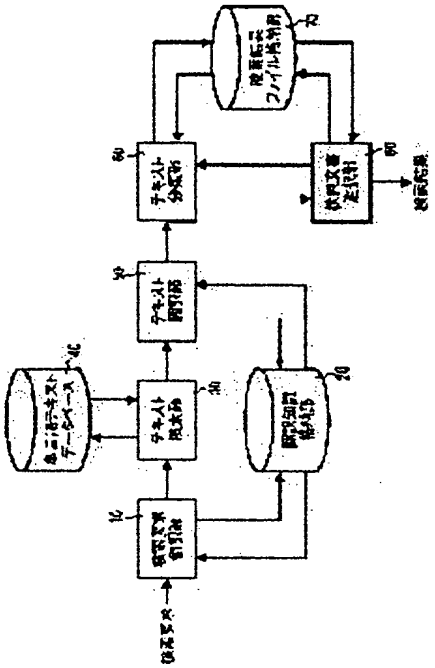
(51)Int.Cl. G06F 17/30
G06F 17/28

(21)Application number : 2001-267671
(22)Date of filing : 04.09.2001

(71)Applicant : JAPAN SCIENCE & TECHNOLOGY CORP
(72)Inventor : ISHIKAWA TETSUYA
FUJII ATSUSHI

(54) MULTI-LINGUAL INFORMATION RETRIEVAL SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a multi-lingual information retrieval system that enables a user to increase in the document selection effectiveness by concurrently retrieving a plurality of multi-lingual databases based on a request from the user and integrating foreign language documents being contained in retrieved results after translating them into the user's native language.
SOLUTION: A multi-lingual information retrieval system translates a retrieval request (a keyword) being inputted by a user in a native language into more than one foreign language and concurrently retrieves native and more than one foreign language documents using the retrieval request. The information retrieval system enables the user to effectively obtain the information using only the user's native language by translating multi-lingual documents contained in the retrieval results into the user's native language, integrating the languages, furthermore, classifying the documents into a plurality of groups based on contents, and displaying the documents that characteristically show each group.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2003－76710
（P2003－76710A）
(43)公開日　平成15年３月14日(2003.3.14)

(51)Int.Cl.[7]		識別記号	FI		テーマコード(参考)
Ｇ０６Ｆ	17/30	３３０	Ｇ０６Ｆ	17/30	３３０Ｃ　５Ｂ０７５
		１７０			１７０Ａ　５Ｂ０９１
	17/28			17/28	Ｘ

審査請求　未請求　請求項の数５　ＯＬ　（全４頁）

(21)出願番号　特願2001－267671(P2001－267671)

(22)出願日　平成13年９月４日(2001.9.4)

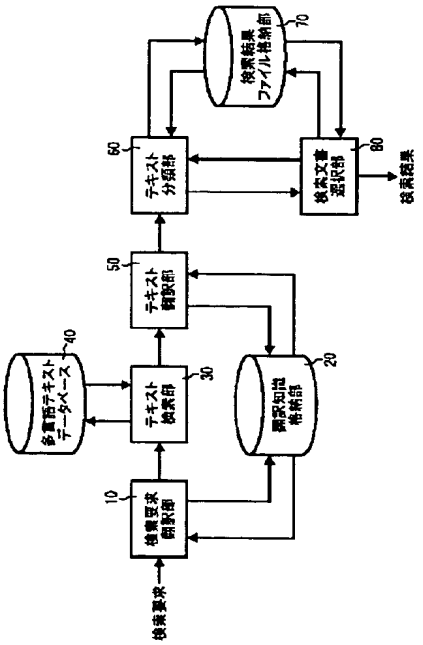
特許法第30条第１項適用申請有り

(71)出願人　396020800
科学技術振興事業団
埼玉県川口市本町４丁目１番８号
(72)発明者　石川　徹也
千葉県松戸市西馬橋４－223
(72)発明者　藤井　敦
茨城県つくば市吾妻１－603－412
(74)代理人　100105371
弁理士　加古　進

Ｆターム(参考)　5B075 ND03 NK02 NR02 NR12 PP22
PQ20 PQ46 QP10 UU06
5B091 AA04 BA03 CD03 CD15

(54)【発明の名称】　多言語情報検索システム

(57)【要約】
【課題】　利用者が入力した検索要求により、複数の言語データベースを同時に検索し、さらに検索結果に含まれる異言語文書を利用者言語に翻訳して統一化した後、利用者の文書選択の効率を向上させる。
【解決手段】　利用者が母国語で入力した検索要求（キーワード）を１種類以上の外国語に翻訳し、それらの検索要求を用いて母国語と１種類以上の外国語の文書情報を同時に検索する。次に検索結果に含まれる異種言語情報を利用者の母国語に翻訳して、言語を統一化し、さらに文書内容に基づいて複数のグループに分類し、各グループを特徴的に表す文書を掲示し、利用者が母国語だけを使って情報を効率よく取得する。



【特許請求の範囲】
【請求項１】多言語情報検索システムであって、
複数の言語のデータベース格納部と、翻訳に用いる知識を格納しておく翻訳知識格納部とを有し、
入力された検索要求を検索対象の文書の言語に翻訳する検索要求翻訳部と、
前記入力された検索要求および前記翻訳された検索要求を基に、前記データベース格納部中から情報を検索する情報検索部と、
前記情報検索部によって検索された情報を特定の言語に統一する情報統一部と、
前記統一された情報を内容により分類する情報分類部と、
該情報分類部により分類された情報を格納しておく検索結果格納部と、
該検索結果格納部より前記分類された情報を、利用者が選択・出力の操作を行う検索情報選択部とを有することを特徴とする多言語情報検索システム。
【請求項２】　請求項１に記載の多言語情報検索システムにおいて、
前記情報選択部は、利用者が選択した情報に対して、前記情報分類部により再度分類を行うことを特徴とする多言語情報検索システム。
【請求項３】　請求項１または２に記載の多言語情報検索システムにおいて、
前記情報分類部は、階層的ベイズ・クラスタリング法を用いて分類することを特徴とする多言語情報検索システム。
【請求項４】　請求項１～３に記載の多言語情報検索システムをコンピュータ・システムに構成させるためのプログラム。
【請求項５】　請求項４に記載の多言語情報検索システムをコンピュータ・システムに構成させるためのプログラムを格納した記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、多言語情報検索に関するものであり、特に、検索結果の選択効率向上に関する。
【０００２】
【技術的背景】ある問題について調べるために、現在は、検索システムを用いてキーワード検索を行い、検索された文書の一覧から欲しい情報を選択して閲覧することが一般的である。しかし、既存の検索システムでは外国語文書の検索は難しいため、キーワードとして入力した利用者の言語のデータベースに、キーワードに関連した文書がない場合は検索に失敗する。また、仮に外国語文書が検索された場合でも、目的とする文書を利用者が選択することは容易ではなく不便である。なお、本明細書において、「母国語」または「利用者言語」とは利用者にとっての母国語であり、「外国語」とは利用者にとっての外国語のことを言う。
【０００３】
【発明が解決しようとする課題】本発明は、利用者が入力した検索要求により、複数の言語データベースを同時に検索し、さらに検索結果に含まれる異言語文書を利用者言語に翻訳して統一化した後、利用者の文書選択の効率を向上させることを目的とする。
【０００４】
【課題を解決するための手段】上記目的を達成するために、本発明は、多言語情報検索システムであって、複数の言語のデータベース格納部と、翻訳に用いる知識を格納しておく翻訳知識格納部とを有し、入力された検索要求を検索対象の文書の言語に翻訳する検索要求翻訳部と、前記入力された検索要求および前記翻訳された検索要求を基に、前記データベース格納部中から情報を検索する情報検索部と、前記情報検索部によって検索された情報を特定の言語に統一する情報統一部と、前記統一された情報を内容により分類する情報分類部と、該情報分類部により分類された情報を格納しておく検索結果格納部と、該検索結果格納部より前記分類された情報を、利用者が選択・出力の操作を行う検索情報選択部とを有する。前記情報選択部は、利用者が選択した情報に対して、前記情報分類部により再度分類を行う。前記情報分類部は、階層的ベイズ・クラスタリング法を用いて分類を行う。本多言語情報検索システムをコンピュータ・システムに構成させるコンピュータ・プログラムおよびコンピュータ・プログラムを記録した記録媒体も本発明である。
【０００５】
【発明の実施の形態】＜本発明の実施形態の概要＞本発明の実施形態は、多言語情報検索システムである。利用者が母国語で入力した検索要求（例えば、キーワード）を１種類以上の外国語に翻訳し、母国語と該１種類以上の外国語の文書情報を同時に検索する。次に検索結果に含まれる異種言語情報を利用者言語に翻訳して、言語を利用者の言語に統一化し、さらに文書内容に基づいて複数のグループに分類し、各グループを特徴的に表す文書を掲示し、利用者が母国語だけを使って情報を効率よく取得するシステムである。
【０００６】＜構成＞図１は本実施形態のシステムの構成を示した図である。以下図面を用いて、本実施形態を詳細に説明する。図１のように本実施形態のシステムは、利用者が入力した検索要求を、翻訳知識を格納する翻訳知識格納部２０を用いて翻訳する検索要求翻訳部１０、利用者の入力した検索要求と翻訳された検索要求を併用して、各言語で書かれた文書類を格納している多言語テキストデータベース４０中の文書を検索するテキスト検索部３０、検索された外国語文書を利用者言語に翻訳するテキスト翻訳部５０、検索文書を内容に基づいて

分類して検索結果ファイルを出力するテキスト分類部６０、検索結果ファイルを格納しておく検索結果ファイル格納部７０、利用者に検索結果ファイル格納部７０中から欲しい文書選択するための検索文書選択部８０で構成している。

【０００７】＜処理の流れ＞

［検索要求の翻訳］利用者が検索要求として単語、句、文章等を利用者の言語により入力する。すると、検索要求翻訳部１０は利用者が入力した検索要求を、辞書や言語モデル等の翻訳知識格納部２０内にある翻訳知識を用いて、検索要求を既存の方法により検索対象のテキストの言語に翻訳する。そして、利用者言語で入力された検索要求と、外国語に翻訳された検索要求を用いて多言語テキストデータベース４０中から文書の検索を行う。

［文書の検索］利用者言語で入力された検索要求と、検索要求翻訳部１０によって外国語に翻訳された検索要求を基に、テキスト検索部３０は多言語テキストデータベース４０を既存の手法を用いて検索し、利用者の検索要求に関連する、各言語で書かれている文書をテキスト翻訳部５０に出力する。

［検索した文書の翻訳］テキスト翻訳部５０は、テキスト検索部３０が出力した検索結果である文書のうち、外国語の文書だけを、検索要求翻訳部１０と同様に、翻訳知識格納部２０内の翻訳知識を用いて利用者言語に翻訳してテキスト分類部６０に出力する。また、検索結果である文書が利用者言語の文書についてはそのままテキスト分類部６０に出力する。また、利用者が検索要求を入力するたびに外国語の文書の翻訳をオンラインで行うだけでなく、オフラインで定期的に翻訳を行ったり、他の利用者に対して行った翻訳結果を再利用することも可能である。この場合は、未訳の文書のみを翻訳すればよいので、処理を効率的に行うことが可能となる。

【０００８】［文書の分類］テキスト分類部６０は、検索された文書の内容に基づいて分類を行う。検索文書を分類する手法としてはいくつかの既知の手法があるが、本発明のシステムの実施形態の１例として階層的ベイズ・クラスタリング（ＨＢＣ：Hierarchical Bayesian Clustering）法を用いる。まず、ＨＢＣ法を説明する。文書集合であるクラスタＣ中の文書をｄとすると、条件付き確率Ｐ（Ｃ｜ｄ）は、文書ｄと文書集合であるクラスタＣとの類似性を示している。この確率Ｐ（Ｃ｜ｄ）を用い、クラスタリングで使う尺度として自己再現率を定義する。あるクラスタＣに関する自己再現率ＳＲ（Ｃ）を以下のように定義する。

【数１】

$$SR(C)=\prod_{d\in C}P(C\mid d)$$

自己再現率は、「クラスタ内の各文書が自分自身を含むクラスタを見つけることができる確率」と解釈することができ、あるクラスタＣにとって、ＳＲ（Ｃ）の値が大きいということは、Ｃ内の各文書を検索入力したとき、それらがＣを見つける確率が高いということである。さて、文書集合Ｄが、クラスタの集合｛C_1，C_2，…｝に分割されているとすると、その文書集合Ｄに対する自己再現率は以下のように定義できる。

【数２】

$$SR(D)=\prod_{C\in D}SR(C)=\prod_{C\in D}\prod_{d\in C}P(C\mid d)$$

これは、文書全体に関する自己検索の精度に関連する。ここまでで、クラスタリングの目的は「文書集合Ｄが与えられたとき、ＳＲ（Ｄ）が最大となる分割を見つけること」と詳細化できる。この文書集合Ｄに対して階層的な二分クラスタ木を構築するために、以下に示す凝集型アルゴリズムを適用する。

①初期クラスタ集合を、Ｄ内の各文書それ自身のみからなるクラスタの集合とする。

②マージによりＳＲ（Ｄ）の増分が最大になるようなクラスタのペアを見つけ、実際にマージする。

③残りのクラスタの数が１でなければ②に戻る。

以上のアルゴリズムをＨＢＣと呼ぶ。なお、このＨＢＣの詳細については、岩山真，徳永健伸，「確率的クラスタリングを用いた文書連想検索」，自然言語処理，Vol.5，No.1，pp.101-118，1998等を参照されたい。このようなＨＢＣ法により、一定数のグループ（クラスタＣ）に分けた検索結果である文書集合Ｄを検索結果ファイル格納部７０に出力することができる。

［検索した文書の選択・出力］利用者は検索結果ファイル格納部７０に分類されて格納された検索・分類結果である文書を、検索文書選択部８０を利用して閲覧しながら、自分の検索要求に関連すると考えられるグループを選択する。そして必要であれば、さらに選択されたグループ内の文書をテキスト分類部６０で再び分類して、利用者に提示することもできる。このように文書分類とグループの選択を再帰的（対話的）に繰り返しながら、利用者が欲しい文書を選択して閲覧できる。

【０００９】上述の実施形態では、検索要求翻訳部１０とテキスト翻訳部５０を別々にしているが、翻訳部として統一して備えることもできる。この場合、検索要求の翻訳と検索結果の文書の翻訳を場合に応じて担当する。上述の検索要求翻訳部１０、テキスト検索部３０、テキスト翻訳部５０、テキスト分類部６０、検索文書選択部８０のそれぞれの処理は、計算機上で実行している。また、その処理を行うためにプログラムを格納した記録媒体から読み出したり、通信回線を介して受信したりしたプログラムを実行する等により、本発明の構成を実現することもできる。この記録媒体には、フロッピー（登録商標）・ディスク、ＣＤ、ＤＶＤ、磁気テープ、ＲＯＭカセット等がある。また、通信回線としては、インター

ネット等がある。

【００１０】

【実施例】本実施例では、多言語テキスト・データベース４０に、日本語と英語の学術抄録のデータセットを用いて行った場合を示す。日本語を母国語とする利用者が検索要求として「パイプライン」というキーワードを入力した。検索要求翻訳部１０にて検索要求「パイプライン」は「pipeline」と訳され、テキスト検索部３０は「パイプライン」と「pipeline」という２つのキーワードを用いて検索を行った。その結果、両者に関連する日英混在の検索結果が得られた。さらに、英語の文書に関してはテキスト翻訳部５０で翻訳される。検索結果である日本語の文書と日本語に翻訳された英語の文書を分類した結果、言語の違いに関係なく、コンピュータ分野で使われる演算パイプライン方式と建築分野の輸送管としてパイプラインの２つのグループに分類された。利用者は興味のあるグループ（分野）を選択して必要な文書だけを母国語で読むことができた。

【００１１】

【発明の効果】本発明によれば、検索要求を母国語で入力しても複数言語文書を検索し、さらに母国語文書と外国語文書が混在していても、言語の違いを意識せずに欲しい文書を容易に取得することができる。また、検索結果である文書を分類して出力するので、利用者が興味のあるグループを選択して必要な文書だけを母国語で読むことができる。

【図面の簡単な説明】

【図１】　本発明の実施形態のシステムの構成を示した図である。

【符号の説明】

１０	検索要求翻訳部
２０	翻訳知識格納部
３０	テキスト検索部
４０	多言語テキスト・データベース
５０	テキスト翻訳部
６０	テキスト分類部
７０	検索結果ファイル格納部
８０	検索文書選択部

【図１】